The Radicalization Risks of GPT-3 and Advanced Neural Language Models

Year: 2020 | Citations: 162 | Authors: Kris McGuffie, Alex Newhouse

Abstract

In this paper, we expand on our previous research of the potential for abuse of generative language models by assessing GPT-3. Experimenting with prompts representative of different types of extremist narrative, structures of social interaction, and radical ideologies, we find that GPT-3 demonstrates significant improvement over its predecessor, GPT-2, in generating extremist texts. We also show GPT-3's strength in generating text that accurately emulates interactive, informational, and influential content that could be utilized for radicalizing individuals into violent far-right extremist ideologies and behaviors. While OpenAI's preventative measures are strong, the possibility of unregulated copycat technology represents significant risk for large-scale online radicalization and recruitment; thus, in the absence of safeguards, successful and efficient weaponization that requires little experimentation is likely. AI stakeholders, the policymaking community, and governments should begin investing as soon as possible in building social norms, public policy, and educational initiatives to preempt an influx of machine-generated disinformation and propaganda. Mitigation will require effective policy and partnerships across industry, government, and civil society.